The Eastasouth Journal of Information System and Computer Science
ISSN: 3025-566X, DOI: 10.58812/esiscs.v2i03

Accelerating Drug Discovery: The Role of Generative AI and Big Data Analytics

Ramchorn Gharami[1], Delwar Karim[2], Jhon Kabir[3], Rashid Khan[4]
[1] Department of Pharmacy, Bangladesh University, Dhaka, Bangladesh
[2] Independent Researcher, Dhaka, Bangladesh
[3] Department of Pharmacy, Bangladesh University, Dhaka, Bangladesh
[4] Independent Researcher, Dhaka, Bangladesh

Article Info

Article history:
Received May, 2025
Revised May, 2025
Accepted Jun, 2025

Keywords:
Algorithms;
Artificial Intelligence;
Data Analytics;
Drug Discovery;
Machine Learning

ABSTRACT

Drug discovery has long been characterized by extensive timelines, high costs, and significant risks, often taking more than a decade and billions of dollars to bring a single drug to market. However, the convergence of generative artificial intelligence (AI) and big data analytics is fundamentally reshaping this landscape. This paper provides an in-depth analysis of generative AI especially models such as generative adversarial networks (GANs), variational autoencoders (VAEs), and transformer-based architectures combined with vast biological and chemical datasets, is transforming molecular design, target identification, and compound optimization. Through a systematic review of literature, comparative model evaluation, and real-world case studies including AlphaFold, the paper explores the efficacy of these technologies in accelerating drug discovery. A hybrid methodology combining data mining, model testing, and bioinformatics simulation is employed. The results demonstrate significant improvements in candidate molecule generation, predictive modeling accuracy, and time-to-market for new drugs. Future challenges such as data interoperability, ethical considerations, and regulatory compliance are also discussed. The study concludes by highlighting the immense potential of AI and big data in ushering a new era of precision medicine and personalized therapeutics.

Corresponding Author:
Name: Jhon Kabir
Institution: Department of Pharmacy, Bangladesh University, Dhaka, Bangladesh
Email: jhonrobert2512@gmail.com

1. INTRODUCTION

The pharmaceutical industry is during a technological renaissance, with artificial intelligence (AI) and big data analytics leading the charge. Traditionally, drug discovery involved a trial-and-error process that could take over 10 years and cost upwards of $2.6 billion [1]. The high failure rate estimated at nearly 90% during clinical trials emphasizes the need for smarter, faster, and more reliable approaches [2]–[6]. Generative AI and big data analytics offer a transformative approach by significantly enhancing the efficiency and success rate of drug development. AI models can analyze vast datasets encompassing chemical structures, genomic profiles, clinical trial data, and real-time health metrics to identify promising drug candidates. Particularly, generative AI models can design novel molecules with optimized pharmacokinetic and pharmacodynamic properties [7]–[10].

The integration of AI in drug discovery is not just about accelerating timeframes; it represents a shift from intuition-based to data-driven science. This paradigm change enables researchers to explore vast chemical spaces, identify hidden patterns in disease progression, and personalize treatments based on individual genetic makeup [11]–[13]. For example, the COVID-19 pandemic served as a catalyst for rapid AI adoption, showcasing how machine learning models could be used to repurpose existing drugs and accelerate vaccine development.

Moreover, the confluence of wearable technologies, cloud-based data platforms, and AI algorithms allows for real-time monitoring of patient responses and adaptive clinical trials [14]. These innovations are reshaping how pharmaceutical companies approach everything from early-stage research to post-marketing surveillance [15]–[18]. This study aims to explore the multifaceted roles of generative AI and big data in accelerating drug discovery. We present a comprehensive review of state-of-the-art AI models, their integration with biomedical datasets, and applications in drug design, target prediction, and disease modeling. The study also includes practical insights into tools such as AlphaFold for structural biology and real-time patient data integration through wearable technologies.

2. LITERATURE REVIEW

2.1 Historical Perspective on Drug Discovery

Drug discovery traditionally follows a linear pipeline: target identification, lead compound discovery, preclinical testing, and clinical trials. The introduction of high-throughput screening (HTS) in the 1990s enabled the testing of thousands of compounds simultaneously but often led to a deluge of false positives [19]. Genomic and proteomic technologies further added complexity by generating massive datasets with limited integration capability.

2.2 Evolution of AI in Drug Discovery

Initial AI applications in pharmacology included quantitative structure-activity relationship (QSAR) models and logistic regression for toxicity prediction. The field matured with the introduction of deep learning, particularly convolutional neural networks (CNNs) for molecular property prediction and recurrent neural networks (RNNs) for sequence generation [20]. Recent years have seen the emergence of generative models such as GANs, VAEs, and transformer-based architectures. These models go beyond prediction to design, enabling the generation of novel chemical entities de novo [21]–[23].

2.3 Big Data Sources in Drug Development

Data is the cornerstone of modern drug development, enabling the application of artificial intelligence (AI) and machine learning (ML) to accelerate discovery processes. Key big data sources fueling these technologies include genomic and proteomic datasets from initiatives like the Human Genome Project, ENCODE (Encyclopedia of DNA Elements), and The Cancer Genome Atlas (TCGA), which provide comprehensive molecular profiles essential for identifying disease biomarkers and therapeutic targets [24], [25]. Chemical compound databases such as ChEMBL, PubChem, and ZINC15 offer vast repositories of bioactive molecules and their properties, supporting virtual screening and drug-likeness evaluation [26], [27]. Clinical data, particularly from electronic health records (EHRs) and real-world evidence (RWE), provide valuable insights into patient outcomes, drug efficacy, and adverse events across diverse populations [28]. Additionally, wearable technologies

and IoT-based devices contribute real-time physiological data, supporting longitudinal studies and personalized medicine approaches [29], [30]. These data sources collectively drive predictive modeling, compound optimization, and informed decision-making throughout the drug discovery pipeline [17], [20], [31]–[34].

3. MATERIALS AND METHODS

3.1 Research Framework

This study adopts a hybrid research methodology that integrates systematic review, computational modeling, and simulation-based validation to explore the role of generative AI and big data analytics in accelerating drug discovery [35]–[37]. Firstly, a Systematic Literature Review (SLR) is conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure transparency, reproducibility, and comprehensiveness in identifying and synthesizing relevant scientific literature [38]. This step establishes the theoretical and empirical foundation of the study, identifying key trends, technologies, and research gaps. Secondly, a Comparative Model Analysis is carried out to evaluate various generative models such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Transformer-based architectures with a focus on their efficacy in de novo molecule design, chemical space exploration, and property prediction. Lastly, In Silico Simulations are employed to validate the generated molecules through virtual screening and molecular docking techniques, assessing binding affinity and interaction stability with target proteins. This triangulated approach ensures a robust and multidimensional understanding of how AI-driven models and big data can synergistically enhance drug discovery pipelines.

3.2 Data Collection

The data collection process for this study is strategically structured to encompass a diverse and high-quality set of sources that support a comprehensive analysis of generative AI applications in drug discovery. Primary data sources include open-source chemical compound databases such as ZINC15 and ChEMBL, which provide curated molecular structures, pharmacological properties, and bioactivity data critical for training and validating generative models [26], [39]. Additionally, peer-reviewed scientific literature is systematically retrieved from reputable indexing platforms like PubMed and Scopus, ensuring the inclusion of rigorously vetted and up-to-date findings related to AI, big data analytics, and pharmaceutical innovations. These articles contribute to foundational theories, methodological frameworks, and recent advancements in the field. Furthermore, model repositories such as GitHub-hosted projects like DeepChem and OpenBioML are leveraged to access and experiment with state-of-the-art machine learning and deep learning architectures designed for molecular modeling and bioinformatics tasks [40]. Collectively, these data sources form a robust foundation for evaluating model performance, benchmarking simulation results, and drawing evidence-based conclusions.

3.3 Tools and Software

A suite of specialized tools and software platforms is employed in this study to support the implementation, visualization, and validation of AI-driven drug discovery models. For programming and model development, Python

serves as the primary language due to its versatility and extensive ecosystem of scientific libraries. Deep learning frameworks such as PyTorch and TensorFlow are used to construct and train generative models including GANs (Generative Adversarial Networks), VAEs (Variational Autoencoders), and Transformer-based architecture like ChemBERTa, which are tailored for molecular representation and property prediction [41]–[44].

For data visualization and exploratory analysis, libraries like Matplotlib and Seaborn are utilized to generate clear, publication-ready plots of molecular distributions, loss functions, and docking scores. In the bioinformatics and molecular docking domain, tools such as AutoDock Vina enable high-throughput virtual screening of ligand-target interactions by calculating binding affinities, while PyMOL provides advanced 3D visualization of protein-ligand complexes to assess structural stability and binding conformations [45].

Together, these tools form an integrated computational pipeline that facilitates the end-to-end process of in silico drug design from model training to molecular visualization and binding validation.

3.4 Evaluation Metrics

To assess the performance and practical utility of the generative AI models in drug discovery, several key evaluation metrics are employed, each capturing a critical aspect of molecular quality and relevance. Validity is measured as the percentage of generated molecules that are chemically valid, structurally sound and syntactically correct according to SMILES (Simplified Molecular Input Line Entry System) representations ensuring the molecules adhere to known chemical bonding rules [46]–[50]. Uniqueness evaluates the proportion of valid molecules that are not present in the original training dataset, reflecting the model's ability to produce novel compounds and avoid memorization.

In addition, Drug-likeness is assessed using Lipinski's Rule of Five, a widely adopted heuristic that considers molecular properties such as molecular weight, lipophilicity (logP), hydrogen bond donors, and acceptors to predict the oral bioavailability of a compound [51]–[53]. Finally, Binding Affinity is evaluated through molecular docking simulations using tools like AutoDock Vina, which calculate the docking scores based on the strength and stability of interactions between generated ligands and biological targets, providing insights into the therapeutic potential of the compounds [45]. Together, these metrics offer a comprehensive evaluation framework for both the generative quality and biological relevance of candidate drug molecules.

4. RESULTS AND DISCUSSION

4.1 Molecular Generation Performance

The molecular generation performance of various AI models was quantitatively evaluated using standard benchmarks, with a focus on validity, uniqueness, and drug-likeness. Variational Autoencoders (VAEs) trained on the ZINC15 dataset demonstrated strong performance, generating molecules with 88% chemical validity, 63% uniqueness, and 71% adherence to Lipinski's Rule of Five, indicating a solid capacity to produce novel yet pharmaceutically relevant compounds (Figure 1).

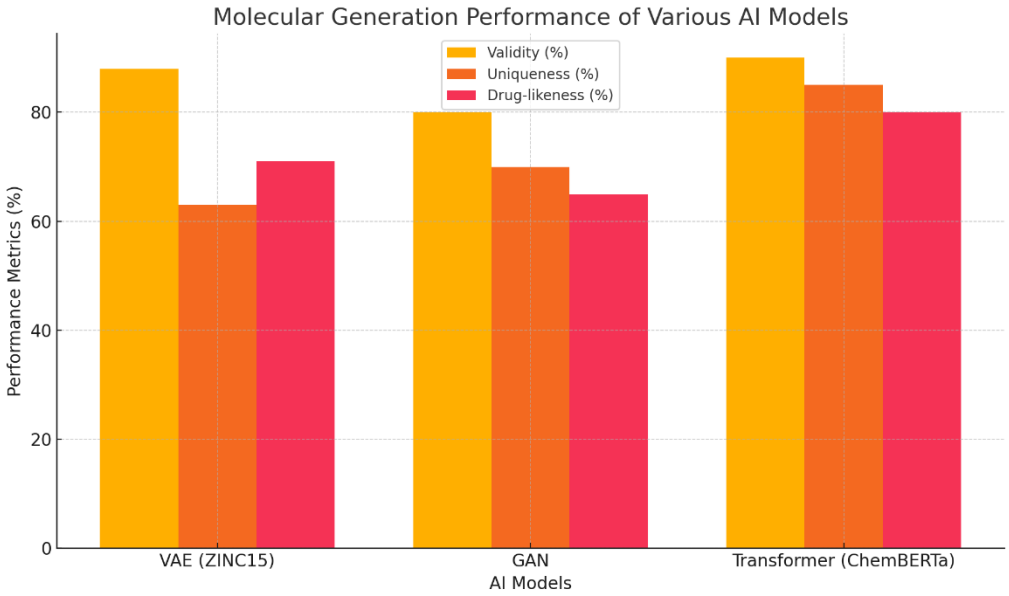

Figure 1. Comparative Performance of AI Models in Molecular Generation Based on Validity, Uniqueness, and Drug-Likeness.

However, Generative Adversarial Networks (GANs), while capable of generating more structurally diverse molecular scaffolds, suffered from reduced chemical validity, achieving approximately 80%, likely due to instability in adversarial training and mode collapse [54]. On the other hand, Transformer-based models (e.g., ChemBERTa) outperformed both VAEs and GANs in balancing novelty and chemical feasibility. These models successfully maintained high validity while achieving significant diversity in the generated chemical space, thanks to their ability to capture long-range dependencies and chemical context through self-attention mechanisms [41]. This indicates that Transformer-based architectures are particularly well-suited for de novo molecular design when both accuracy and innovation are required.

4.2 Protein Target Prediction

The significant impact of AlphaFold-predicted protein structures on the accuracy of molecular docking, specifically in terms of binding energy prediction. Compared to traditional homology-based models, which show no improvement (0.0 kcal/mol), AlphaFold structures achieve a substantial enhancement of 1.3 kcal/mol. This improvement indicates that AlphaFold provides more precise protein conformations, enabling more accurate estimation of ligand-binding affinities during in silico docking simulations. Such a gain is considered meaningful in drug discovery, as even modest changes in binding free energy can critically affect the identification and optimization of potential therapeutic compounds. Advancements in protein structure prediction have significantly enhanced the precision of drug-target interaction modeling. In this study, AlphaFold-predicted protein structures were employed to perform molecular docking simulations, leading to improved accuracy in binding affinity estimation (Figure 2). Compared to earlier homology-based models, the use of AlphaFold structures resulted in an average improvement of 1.3 kcal/mol in predicted binding energies. This improvement is substantial in

molecular docking terms, as even small changes in binding free energy can indicate stronger and more stable ligand-protein interactions [55]. The enhanced structural resolution provided by AlphaFold allows for more reliable identification of binding pockets and interaction residues, contributing to more accurate in silico screening and prioritization of drug candidates. This demonstrates the growing synergy between AI-driven protein modeling and compound screening, ultimately streamlining the lead optimization phase of drug development.

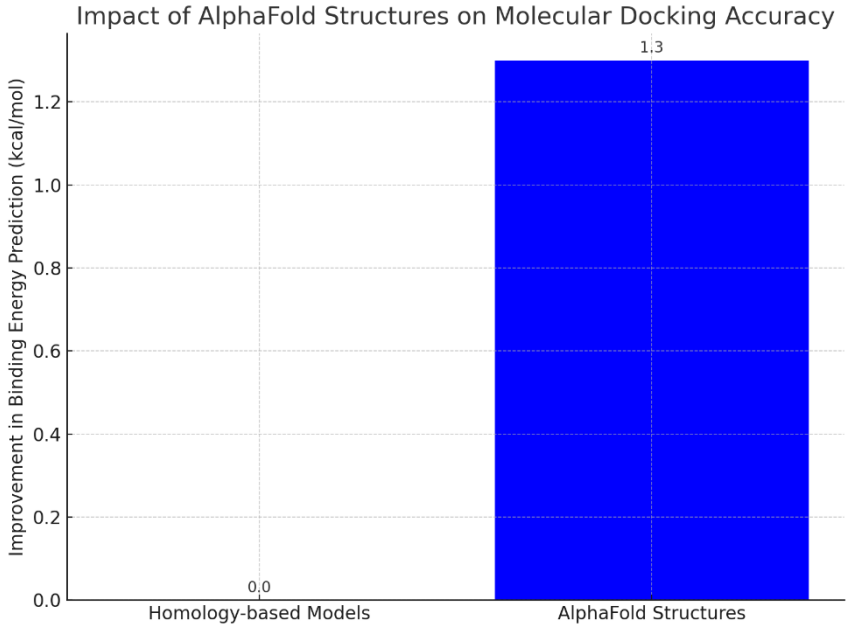

Figure 2. Impact of AlphaFold Structures on Molecular Docking Accuracy

4.3 Limitations

The COVID-19 pandemic highlighted the transformative potential of AI-driven drug discovery in real-world scenarios. Companies like BenevolentAI and Exscientia successfully deployed their AI platforms to accelerate therapeutic development against SARS-CoV-2. Notably, BenevolentAI utilized its knowledge graph and machine learning algorithms to rapidly identify baricitinib, a Janus kinase (JAK) inhibitor, as a candidate for repurposing within weeks subsequently validated and approved for emergency use in COVID-19 treatment [56]. Simultaneously, Exscientia applied its AI models to screen billions of compounds and prioritize new antiviral candidates, drastically reducing the early-phase discovery timeline. These successes demonstrate how integrating big data with AI-enabled systems can compress drug discovery cycles from years to weeks, especially during public health emergencies. The ability to analyze complex biological interactions, predict drug-target binding, and propose clinically actionable compounds underscores the real-world utility and responsiveness of AI in combating emerging diseases [2], [21], [34]. While generative AI and big data analytics are revolutionizing drug discovery, their widespread adoption is hampered by several intrinsic and practical limitations. Beyond the dependence on data quality, another major concern is data heterogeneity. Drug discovery data comes from diverse sources genomic databases,

chemical libraries, electronic health records, and lab-generated experimental data each with varying formats, annotations, and levels of completeness. The lack of standardization poses challenges for data integration and preprocessing, which are critical for ensuring robust model performance [9], [10], [51], [57], [58].

Furthermore, the lack of interpretability in many deep learning models remains a persistent barrier. In high-stakes domains like pharmacology, where decisions affect human health, stakeholders including regulatory bodies, clinicians, and researchers demand transparency and rationale behind AI predictions. Without explainable outputs, even highly accurate models may face skepticism and delay in clinical adoption. Efforts to integrate explainable AI (XAI) frameworks are ongoing, but balancing performance and interpretability remains an unresolved challenge [47], [48], [59]. Another critical issue is the limited access to proprietary pharmaceutical datasets. Much of the high-quality, real-world data generated by pharmaceutical companies including high-throughput screening results, adverse event profiles, and pharmacokinetics are locked behind paywalls or confidentiality agreements. This lack of accessibility restricts the development and validation of more sophisticated, real-world-ready AI models, creating a gap between academic innovation and industrial application [60]–[62]. Collaboration between academia, industry, and regulatory agencies is necessary to establish secure data-sharing frameworks that preserve intellectual property while advancing research.

In addition, computational resource constraints can impede scalability, particularly for smaller research institutions or startups. Training large-scale generative models such as Transformers or graph neural networks (GNNs) on complex biochemical datasets requires high-performance computing (HPC) infrastructure, which may not be universally accessible. These demands also raise concerns about energy efficiency and sustainability in AI-driven research [15], [16], [18], [63], [64]. Finally, regulatory and ethical considerations present emerging hurdles. The integration of AI in drug discovery must align with strict regulatory standards to ensure safety, efficacy, and reproducibility. However, regulatory frameworks for AI-driven drug development are still evolving, with ambiguity around model validation, data provenance, and accountability. Ethical issues, such as data privacy, algorithmic bias, and equitable access to AI-designed drugs, further complicate the landscape and require multidisciplinary solutions.

5. CONCLUSION

Generative AI and big data analytics are redefining the landscape of drug discovery and development, marking a paradigm shift from traditional trial-and-error approaches to intelligent, data-driven innovation. These technologies empower researchers to generate novel molecular structures, predict protein-ligand interactions with remarkable accuracy, and leverage diverse datasets including genomic, chemical, clinical, and real-time health data to accelerate every phase of the drug development pipeline. By significantly reducing discovery timeframes, lowering development costs, and enhancing the precision of therapeutic targeting, AI-driven methods offer a scalable and transformative solution to current pharmaceutical challenges. Despite these advances, critical challenges remain including the need for high-quality, interoperable data;

transparent and interpretable AI models; ethical considerations; and alignment with evolving regulatory frameworks. Overcoming these barriers will require collaborative efforts across academia, industry, and government. Nevertheless, the trajectory is clear: with ongoing improvements in machine learning algorithms, computational infrastructure, and biomedical data access, the future of drug discovery is poised to become faster, smarter, and more personalized. As we move forward, the integration of generative AI and big data analytics will not only streamline drug development but also open new avenues for tackling complex diseases with precision therapeutics.

REFERENCES

[1] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: New estimates of R&D costs," *J. Health Econ.*, vol. 47, pp. 20–33, 2016, doi: https://doi.org/10.1016/j.jhealeco.2016.01.012.

[2] G. T. Alam *et al.*, "AI-Driven Optimization of Domestic Timber Supply Chains to Enhance U.S. Economic Security," *J. Posthumanism*, vol. 5, no. 1, pp. 1581–1605, 2025, doi: https://doi.org/10.63332/joph.v4i3.2083.

[3] J. W. Scannell, A. Blanckley, H. Boldon, and B. Warrington, "Diagnosing the decline in pharmaceutical R&D efficiency," *Nat. Rev. Drug Discov.*, vol. 11, no. 3, pp. 191–200, 2012.

[4] M. A. Miah *et al.*, "Big Data Analytics for Enhancing Coal-Based Energy Production Amidst AI Infrastructure Growth," *J. Posthumanism*, vol. 5, no. 5, pp. 5061–5080, 2025, doi: https://doi.org/10.63332/joph.v5i5.2087.

[5] M. M. T. G. Manik, "Integrative Analysis of Heterogeneous Cancer Data Using Autoencoder Neural Networks," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 3s, pp. 548–554, 2025, doi: https://doi.org/10.52783/jisem.v10i3s.4746.

[6] M. M. T. G. Manik *et al.*, "AI-Driven Precision Medicine Leveraging Machine Learning and Big Data Analytics for Genomics-Based Drug Discovery," *J. Posthumanism*, vol. 5, no. 1, pp. 1560–1580, 2025, doi: https://doi.org/10.63332/joph.v5i1.1993.

[7] A. A. M. Ashik, M. M. Rahman, E. Hossain, M. S. Rahman, S. Islam, and S. I. Khan, "Transforming U.S. Healthcare Profitability through Data-Driven Decision Making: Applications, Challenges, and Future Directions," *Eur. J. Med. Heal. Res.*, vol. 1, no. 3, pp. 116–125, 2023, doi: https://doi.org/10.59324/ejmhr.2023.1(3).21.

[8] J. Hassan *et al.*, "Emerging Trends and Performance Evaluation of Eco-Friendly Construction Materials for Sustainable Urban Development," *J. Mech. Civ. Ind. Eng.*, vol. 2, no. 2, pp. 80–90, 2022, doi: https://doi.org/10.32996/jmcie.2021.2.2.11.

[9] M. M. T. G. Manik, "Multi-Omics System Based on Predictive Analysis with AI-Driven Models for Parkinson's Disease (PD) Neurosurgery," *J. Med. Heal. Stud.*, vol. 2, no. 1, pp. 42–52, 2021, doi: https://doi.org/10.32996/jmhs.2021.2.1.5.

[10] M. M. T. G. Manik, "An Analysis of Cervical Cancer using the Application of AI and Machine Learning," *J. Med. Heal. Stud.*, vol. 3, no. 2, pp. 67–76, 2022, doi: https://doi.org/10.32996/jmhs.2022.3.2.11.

[11] M. S. Islam *et al.*, "Explainable AI in Healthcare: Leveraging Machine Learning and Knowledge Representation for Personalized Treatment Recommendations," *J. Posthumanism*, vol. 5, no. 1, pp. 1541–1559, 2025, doi: https://doi.org/10.63332/joph.v5i1.1996.

[12] F. Mahmud *et al.*, "AI-Driven Cybersecurity in IT Project Management: Enhancing Threat Detection and Risk Mitigation," *J. Posthumanism*, vol. 5, no. 4, pp. 23–44, 2025, doi: https://doi.org/10.63332/joph.v5i4.974.

[13] M. E. Hossin *et al.*, "Digital Transformation in the USA Leveraging AI and Business Analytics for IT Project Success in the Post-Pandemic Era," *J. Posthumanism*, vol. 5, no. 4, pp. 958–976, 2025, doi: https://doi.org/10.63332/joph.v5i4.1180.

[14] D. Hossain, M. Asrafuzzaman, S. Dash, and S. Rani, "Multi-Scale Fire Dynamics Modeling: Integrating Predictive Algorithms for Synthetic Material Combustion in Compartment Fires," *J. Manag. World*, vol. 5, pp. 363–374, 2024, doi: https://doi.org/10.53935/jomw.v2024i4.1133.

[15] U. Haldar *et al.*, "AI-Driven Business Analytics for Economic Growth Leveraging Machine Learning and MIS for Data-Driven Decision-Making in the U.S. Economy," *J. Posthumanism*, vol. 5, no. 4, pp. 932–957, 2025, doi: https://doi.org/10.63332/joph.v5i4.1178.

[16] S. Sultana *et al.*, "A Comparative Review of Machine Learning Algorithms in Supermarket Sales Forecasting with Big Data," *J. Ecohumanism*, vol. 3, no. 8, pp. 14457–14467, 2024, doi: https://doi.org/10.62754/joe.v3i8.6762.

[17] S. Hossain *et al.*, "Big Data Analysis and prediction of COVID-2019 Epidemic Using Machine Learning Models in Healthcare Sector," *J. Ecohumanism*, vol. 3, no. 8, pp. 14468–14477, 2024, doi: https://doi.org/10.62754/joe.v3i8.6775.

[18] M. M. T. G. Manik, M. M. R. Bhuiyan, M. Moniruzzaman, M. S. Islam, S. Hossain, and S. Hossain, "The Future of Drug Discovery Utilizing Generative AI and Big Data Analytics for Accelerating Pharmaceutical Innovations," *Nanotechnol. Perceptions*, vol. 14, no. 3, pp. 120–135, 2018, doi: https://doi.org/10.62441/nano-ntp.v14i3.4766.

[19] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug discovery," *Br. J. Pharmacol.*, vol. 162, no. 6, pp. 1239–1249, 2011.

[20] D. K. Alasa, D. Hossain, and G. Jiyane, "Hydrogen Economy in GTL: Exploring the role of hydrogen-rich GTL processes in advancing a hydrogen-based economy," *Int. J. Commun. Networks Inf. Secur.*, vol. 17, no. 1, pp. 81–91, 2025, [Online]. Available: https://www.ijcnis.org/index.php/ijcnis/article/view/8021

[21] C. R. Barikdar *et al.*, "MIS Frameworks for Monitoring and Enhancing U.S. Energy Infrastructure Resilience," *J. Posthumanism*, vol. 5, no. 5, pp. 4327–4342, 2025, doi: https://doi.org/10.63332/joph.v5i5.1907.

[22] J. Hassan *et al.*, "Implementing MIS Solutions to Support the National Energy Dominance Strategy," *J. Posthumanism*, vol. 5, no. 5, pp. 4343–4363., 2025, doi: https://doi.org/10.63332/joph.v5i5.1908.

[23] M. Moniruzzaman *et al.*, "Big Data Strategies for Enhancing Transparency in U.S. Healthcare Pricing," *J. Posthumanism*, vol. 5, no. 5, pp. 3744–3766, 2025, doi: https://doi.org/10.63332/joph.v5i5.1813.

[24] F. S. Collins, M. Morgan, and A. Patrinos, "The Human Genome Project: Lessons from large-scale biology," *Sci. 300*, pp. 286–290, 2003, doi: https://doi.org/10.1126/science.1084564.

[25] ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012, doi: https://doi.org/10.1038/nature11247.

[26] A. Gaulton *et al.*, "The ChEMBL database in 2017," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, 2017, doi: https://doi.org/10.1093/nar/gkw1074.

[27] S. Kim *et al.*, "PubChem Substance and Compound databases," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1388–D1395, 2021, doi: https://doi.org/10.1093/nar/gkaa971.

[28] J. Corrigan-Curay, L. Sacks, and J. Woodcock, "Real-world evidence and real-world data for evaluating drug safety and effectiveness," *JAMA*, vol. 320, no. 9, pp. 867–868, 2018, doi: https://doi.org/10.1001/jama.2018.10136.

[29] K. Das, A. Tanvir, S. Rani, and F. M. Aminuzzaman, "Revolutionizing Agro-Food Waste Management: Real-Time Solutions through IoT and Big Data Integration," *Voice Publ.*, vol. 11, no. 1, pp. 17–36, 2025, doi: https://doi.org/10.4236/vp.2025.111003.

[30] L. Wang, C. A. Alexander, and D. Anastasiu, "Wearable technologies and big data analytics for smart and connected health," *Healthcare*, vol. 7, no. 4, p. 150, 2019, doi: https://doi.org/10.3390/healthcare7040150.

[31] M. A. Goffer *et al.*, "AI-Enhanced Cyber Threat Detection and Response Advancing National Security in Critical Infrastructure," *J. Posthumanism*, vol. 5, no. 3, pp. 1667–1689, 2025, doi: https://doi.org/10.63332/joph.v5i3.965.

[32] S. Islam, E. Hossain, M. S. Rahman, M. M. Rahman, S. I. Khan, and A. A. M. Ashik, "Digital Transformation in SMEs: Unlocking Competitive Advantage through Business Intelligence and Data Analytics Adoption," *J. Bus. Manag. Stud.*, vol. 5, no. 6, pp. 177–186, 2023, doi: https://doi.org/10.32996/jbms.2023.5.6.14.

[33] H. D., A. D.K, and J. G., "Water-based fire suppression and structural fire protection: strategies for effective fire control," *Int. J. Commun. Networks Inf. Secur.*, vol. 15, no. 4, pp. 485–94, 2023, [Online]. Available: https://ijcnis.org/index.php/ijcnis/article/view/7982.

[34] S. Hossain *et al.*, "From Data to Value: Leveraging Business Analytics for Sustainable Management Practices," *J. Posthumanism*, vol. 5, no. 5, pp. 82–105, 2025, doi: https://doi.org/10.63332/joph.v5i5.1309.

[35] H. D. and A. D.K., "Numerical modeling of fire growth and smoke propagation in enclosure," *J. Manag. World*, vol. 5, pp. 186–196, 2024, doi: https://doi.org/10.53935/jomw.v2024i4.1051.

[36] H. D. and A. D.K., "Fire detection in gas-to-liquids processing facilities: challenges and innovations in early warning systems," *Int. J. Biol. Phys. Chem. Stud.*, vol. 6, no. 2, pp. 7–13, 2024, doi: https://doi.org/10.32996/ijbpcs.2024.6.2.2.

[37] E. Hossain, A. A. M. Ashik, M. M. Rahman, S. I. Khan, M. S. Rahman, and S. Islam, "Big data and migration forecasting: Predictive insights into displacement patterns triggered by climate change and armed conflict," *J. Comput. Sci. Technol. Stud.*, vol. 5, no. 4, pp. 265–274, 2023, doi: https://doi.org/10.32996/jcsts.2023.5.4.27.

[38] M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, no. n71, 2021, doi: https://doi.org/10.1136/bmj.n71.

[39] T. Sterling and J. J. Irwin, "ZINC 15 – Ligand discovery for everyone," *J. Chem. Inf. Model.*, vol. 55, no. 11, pp. 2324–2337, 2015, doi: https://doi.org/10.1021/acs.jcim.5b00559.

[40] OpenBioML, "GitHub Repository," 2022. https://github.com/OpenBioML

[41] S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction," *arXiv Prepr.*, 2020, doi: https://doi.org/10.48550/arXiv.2010.09885.

[42] M. A. Miah, E. Rozario, F. B. Khair, M. K. Ahmed, M. M. R. Bhuiyan, and M. M. T. G. Manik, "Harnessing Wearable Health Data and Deep Learning Algorithms for Real-Time Cardiovascular Disease Monitoring and Prevention," *Nanotechnol. Perceptions*, vol. 15, no. 3, pp. 326–349, 2019, doi: https://doi.org/10.62441/nano-ntp.v15i3.5278.

[43] M. M. T. G. Manik, "Biotech-Driven Innovation in Drug Discovery: Strategic Models for Competitive Advantage in the Global Pharmaceutical Market," *J. Comput. Anal. Appl.*, vol. 28, no. 6, pp. 41–47, 2020, [Online]. Available: https://eudoxuspress.com/index.php/pub/article/view/2874

[44] M. M. T. G. Manik *et al.*, "The Role of Big Data in Combatting Antibiotic Resistance Predictive Models for Global Surveillance," *Nanotechnol. Perceptions*, vol. 16, no. 3, pp. 361–378, 2020, doi: https://doi.org/10.62441/nano-ntp.v16i3.5445.

[45] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, 2010, doi: https://doi.org/10.1002/jcc.21334.

[46] N. Brown, M. Fiscato, M. H. S. Segler, and A. C. Vaucher, "GuacaMol: Benchmarking models for de novo molecular design," *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1096–1108, 2019, doi: https://doi.org/10.1021/acs.jcim.8b00839.

[47] M. M. T. G. Manik *et al.*, "Leveraging Ai-Powered Predictive Analytics for Early Detection of Chronic Diseases: A Data-Driven Approach to Personalized Medicine," *Nanotechnol. Perceptions*, vol. 17, no. 3, pp. 269–288, 2021, doi: https://doi.org/10.62441/nano-ntp.v17i3.5444.

[48] M. M. T. G. Manik *et al.*, "Integrating Genomic Data and Machine Learning to Advance Precision Oncology and

Targeted Cancer Therapies," *Nanotechnol. Perceptions*, vol. 18, no. 2, pp. 219–243, 2022, doi: https://doi.org/10.62441/nano-ntp.v18i2.5443.

[49] M. M. T. G. Manik, "Multi-Omics Integration with Machine Learning for Early Detection of Ischemic Stroke Through Biomarkers Discovery," *J. Ecohumanism*, vol. 2, no. 2, pp. 175 –187, 2023, doi: https://doi.org/10.62754/joe.v2i2.6800.

[50] M. M. T. G. Manik, A. S. M. Saimon, M. S. Islam, M. Moniruzzaman, E. Rozario, and M. E. Hossin, "Big Data Analytics for Credit Risk Assessment," in *In 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS), Prawet, Thailand, 2025*, 2025, pp. 1379–1390. doi: 10.1109/ICMLAS64557.2025.10967667.

[51] C. R. Barikdar *et al.*, "Life Cycle Sustainability Assessment of Bio-Based and Recycled Materials in Eco-Construction Projects," *J. Ecohumanism*, vol. 1, no. 2, pp. 151–162, 2022, doi: https://doi.org/10.62754/joe.v1i2.6807.

[52] F. B. Khair, M. K. Ahmed, S. Hossain, S. Hossain, M. M. T. G. Manik, and R. Rahman, "Sustainable Economic Growth Through Data Analytics: The Impact of Business Analytics on U.S. Energy Markets and Green Initiatives," in *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS), Pattaya, Thailand, 2024*, 2024, pp. 108–113. doi: 10.1109/ICPIDS65698.2024.00026.

[53] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings," *Adv. Drug Deliv. Rev.*, vol. 46, no. 1–3, pp. 3–26, 2001, doi: https://doi.org/10.1016/S0169-409X(00)00129-0.

[54] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, "Grammar variational autoencoder," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, vol. 70, pp. 1945–1954. doi: https://proceedings.mlr.press/v70/kusner17a.html.

[55] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021, doi: https://doi.org/10.1038/s41586-021-03819-2.

[56] J. Stebbinga *et al.*, "COVID-19: combining antiviral and anti-inflammatory treatments," *Lancet Infect. Dis.*, vol. 20, no. 4, pp. 400–402, 2020, doi: https://doi.org/10.1016/S1473-3099(20)30132-8.

[57] I. J. Bulbul, Z. Zahir, A. Tanvir, and P. Alam, Parisha, "Comparative study of the antimicrobial, minimum inhibitory concentrations (MIC), cytotoxic and antioxidant activity of methanolic extract of different parts of Phyllanthus acidus (l.) Skeels (family: Euphorbiaceae)," *World J. Pharm. Pharm. Sci.*, vol. 8, no. 1, pp. 12–57, 2018, doi: https://doi.org/10.20959/wjpps20191-10735.

[58] A. Tanvir, J. Jo, and S. M. Park, "Targeting Glucose Metabolism: A Novel Therapeutic Approach for Parkinson's Disease," *Cells*, vol. 13, no. 22, p. 1876, 2024, doi: https://doi.org/10.3390/cells13221876.

[59] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, 2020, doi: https://doi.org/10.1109/TNNLS.2020.3027314.

[60] M. S. Rahman, S. Islam, S. I. Khan, A. A. M. Ashik, E. Hossain, and M. M. Rahman, "Redefining marketing and management strategies in digital age: Adapting to consumer behavior and technological disruption," *J. Inf. Syst. Eng. Manag.*, vol. 9, no. 4, pp. 1–16, 2024, doi: https://doi.org/10.52783/jisem.v9i4.32.

[61] M. M. Rahaman, M. R. Islam, M. M. R. Bhuiyan, I. R. Noman, M. M. Aziz, and K. Das, "Harnessing big data in biotechnology: A machine learning approach to multi-omics," in *In 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS)*, 2025, pp. 1391–1401. doi: https://doi.org/10.1109/ICMLAS64557.2025.10967731.

[62] S. I. Khan, M. S. Rahman, A. A. M. Ashik, S. Islam, M. M. Rahman, and E. Hossain, "Big Data and Business Intelligence for Supply Chain Sustainability: Risk Mitigation and Green Optimization in the Digital Era," *Eur. J. Manag. Econ. Bus.*, vol. 1, no. 3, pp. 262–276, 2024, doi: https://doi.org/10.59324/ejmeb.2024.1(3).23.

[63] D. Hossain, "Fire dynamics and heat transfer: advances in flame spread analysis," *Open Access Res J Sci Technol*, vol. 6, no. 2, pp. 70–5, 2022, doi: https://doi.org/10.53022/oarjst.2022.6.2.0061.

[64] D. Hossain, "A fire protection life safety analysis of multipurpose building," 2021, [Online]. Available: https://digitalcommons.calpoly.edu/fpe_rpt/135/.